EAGER: High-throughput, culture-independent technique identifying cyanobacteria infections to improve understanding of carbon biogeochemical cycling

A Data management plan created using the DMPTool

Creator(s): Sarah Preheim

Affiliation: Johns Hopkins University

Last modified: December 07, 2017

EAGER: High-throughput, culture-independent technique identifying cyanobacteria infections to improve understanding of carbon biogeochemical cycling

---

Data Policy Compliance

The project investigators will comply with the data management and dissemination policies described in the *NSF Award and Administration Guide* (AAG, Chapter VI.D.4) and the *NSF Division of Ocean Sciences Sample and Data Policy*.

Pre-Cruise Planning

Not applicable.

Description of Data Types

The project will produce several observational and experimental datasets, described below. In addition to the datasets described below, computational and educational resources produced by the project, including scripts, data and images, will be made available for public use through the PI's Github account and on the Preheim Lab website (http://preheimlab.johnshopkins.edu/). Observational data will be collected from the Chesapeake Bay from March 2018-March 2019.

Observational Datasets:

1. *Environmental measurements:* Environmental data will be collected using a YSI EXO1 sonde; data will include standard environmental measurements such as pressure, temperature, salinity, oxygen, turbidity, ORP, pH and conductivity. Chlorophyll and FDOM measurements collected at the SERC sample site by the Smithsonian National Museum of Natural History SI Physical Monitoring Network (nmnhmp.riocean.com) will also be recorded. File types: Excel spreadsheet and ASCII files. Repository: BCO-DMO.
2. *Sampling event log:* Sampling event log; will include unique sample numbers, start/end dates, start/end times, GPS coordinates of locations, replicate number, weather information, tidal events, and notes about sampling. Sampling event logs will be recorded on paper log sheets or with a computer. File types: Excel file converted to .csv; scanned PDFs. Repository: BCO-DMO

Experimental Datasets:

1. *Microscope Images:* Images collected from fluorescence microscopy will be stored on the Maryland Advanced Research Computing Center (MARCC) as .jpeg or .tiff images and backed up to a server. An image inventory will be created associating the images with the samples. Repository: BCO-DMO
2. *Plaque assay results*: Cyanobacteria isolates, infecting phage and infection networks will be generated from plaque assays. Strains will be archived as described below. Infection results will be stored in an excel spreadsheet. Repository: BCO-DMO.
3. *Genetic sequencing:* DNA sequences from samples collected at the sampling site will include 16S rRNA gene and viral marker gene amplicon data, as well as shotgun metagenomic sequencing of various components

of the microbial community, including the viral and bacterial communities. EpicPCR data of associated virus and host amplicons will also be generated. 16S rRNA gene and viral marker gene data will be sequenced from isolates and plaques during plaque assays. Sequencing will be performed at the JHU Genetic Resource Core Facility, Deep Sequencing and Microarray Core Facility or University of Maryland Institute for Genome Sciences. File types: Quality (.fastq) and sequence data (.fasta) files. Repository: NCBI Short-read archive (SRA); accession numbers to be provided to BCO-DMO.

Data and Metadata Formats and Standards

Field observation data will be stored in flat ASCII files, which can be read easily by different software packages. Field data will include date, time, latitude, longitude and depth, as appropriate. Quality flags will be assigned according to the ODS IODE Quality Flag scheme when appropriate (IOC Manuals and Guides, 54, volume 3; http://www.iode.org/mg54_3). Metadata will be prepared in accordance with BCO-DMO conventions (i.e. using the BCO-DMO metadata forms) and will include detailed descriptions of collection and analysis procedures.

Standard operating procedures (SOP) are in place to ensure sampling quality during preparation of the sampling equipment, instrument calibration, sample collection and storage. SOPs are in place for laboratory work for DNA extraction, diversity surveys and sample sequencing. Positive and negative controls are required for each project, including field blanks, DNA extraction negatives, positive and negative polymerase chain reaction (PCR) controls and replicates. Samples are processed randomly to minimize the impact of batch effects. SOPs are in place for sequence analysis. All sequence data undergoes stringent quality filtering prior to analysis. Standard pipelines will be comprised of all of the processing steps in detail, including the script used to process the data.

Data Storage and Access During the Project

The investigators will store project data (including sequences data, spreadsheets, ASCII files, images, and PDFs of scanned logs) on personal computers that are backed up by the Johns Hopkins University's central IT organization through CrashPlan to an WSE server and an off-site Cloud server. Additionally, computers will also have personal backups, such as Apple Time Machine to an onsite external hard drive. The Principal Investigator (PI) has also established an account with Maryland Advanced Research Computing Center of for data storage, computing and sharing among project investigators.

Mechanisms and Policies for Access, Sharing, Re-Use, and Re-Distribution

DNA sequences will be deposited in the National Center for Biotechnology Information (NCBI) database GenBank upon submission of manuscripts, but no more than 2 years after sample collection. GenBank accession numbers will be provided to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) in an Excel spreadsheet or .CSV file and metadata will be provided using the BCO-DMO Dataset Metadata submission form. Data sets produced will be made available through the BCO-DMO data system within two-years from the date of collection. Scripts used to process epicPCR data (bash and/or python scripts) will be made publically available through the PI's Github account and the Preheim Lab website. The

project investigators will work with BCO-DMO data managers to make project data available online in compliance with the NSF OCE Sample and Data Policy. Data, samples, and other information collected under this project can be made publically available without restriction once submitted to the public repositories. Data produced by this project may be of interest to biological oceanographers interested in the influence of viruses on microbial communities. We will adhere to and promote the standards, policies, and provisions for data and metadata submission, access, re-use, distribution, and ownership as prescribed by the BCO-DMO Terms of Use (http://www.bco-dmo.org/terms-use).

Plans for Archiving

BCO-DMO will also ensure that project data are submitted to the appropriate national data archive. The PI will work with BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data. Characterized cyanobacteria and phage will be archived in glycerol stocks for 5 years after publication of our results at -80 °C. We will work with with various culture collections (Provasoli-Guillard National Center for Marine Algae and Microbiota and Culture Collection of Algae at the University of Texas) to archive strains and associated phage for distribution to the public. DNA, gylcerols and unfiltered water will be preserved for 5 years after publication of any manuscript.

Roles and Responsibilities

Dr. Preheim will be responsible for sharing data among the project participants in a timely fashion. Dr. Sakowski will be responsible for sample collection, molecular biology, sequencing, culturing and will submit the resulting sequences to the National Center for Biotechnology Information's (NCBI) GenBank database. Dr. Preheim will coordinate the overall data management and data sharing and will submit the project data, including GenBank accession numbers, and metadata to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) who will be responsible for forwarding these data and metadata to the appropriate national archive.